机器学习与知识发现 提纲

Ch2.模型评估与选择

定义：错误率，精度，误差，经验误差，泛化误差，过拟合，欠拟合 P23

评估方法：留出法 P25

k折交叉验证 P26

自助法：对数据集有放回采样得到训练集，其余测试集 P27

性能度量：

回归任务-均方误差： $E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(\boldsymbol{x}_i)-y_i)^2$ P29

分类任务-错误率&精度： $E(f;D)=1-acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$ P29

查准率&查全率：混淆矩阵，$Precision=\frac{TP}{TP+FP}, Recall=\frac{TP}{TP+FN}$ P30

P-R曲线（查准率-查全率曲线）：平衡点(BEP): 查准率=查全率（P-R曲线与$y=x$连线) P31

F1度量：$F1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{\text{总数}+TP-TN}$ P32

$F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}$,$\beta>1$,偏重查全率；$\beta<1$,偏重查准率

多个二分类混淆矩阵：macro&micro, ppt没讲，P32-33有

ROC: 纵轴：真正例率 $TPR=\frac{TP}{TP+FN}$ 横轴：假正例率 $FPR=\frac{FP}{TN+FP}$

绘制过程参照P34 AUC=ROC曲线下面积 若ROC曲线由有限个点按序连接，则可估算

$AUC=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$ P34-35

代价敏感错误率：权衡不同类型错误造成的不同损失，定义$cost_{01},cost_{10}$,代价敏感错误率为

$$E(f;D;cost)=\frac{1}{m}(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{01}+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i\neq y_i)\times cost_{10})$$

代价曲线：横轴 $P(+)cost=\frac{p\times cost_{01}}{p\times cost_{01}+(1-p)\times cost_{10}}$,

纵轴 $cost_{norm}=\frac{FNR\times p\times cost_{01}+FPR\times(1-p)\times cost_{10}}{p\times cost_{01}+(1-p)\times cost_{10}}$ P36

ROC曲线上每个点对应代价曲线上一条线段，所有线段的下界围成学习器的期望总体代价。

比较检验： 二项检验，t检验，交叉验证t检验，McNemar检验，Friedman检验，Nemenyi后续检验，

大概率不考？P38-44

偏差与方差： $var(\boldsymbol{x})=\mathbb{E}_D[(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))^2]$, $bias^2(\boldsymbol{x})=(\bar{f}(\boldsymbol{x})-y)^2$

有泛化误差$E(f;D)=bias^2(\boldsymbol{x})+var(\boldsymbol{x})+\varepsilon^2$

偏差：算法本身拟合能力 方差：数据扰动造成的影响 噪声：学习问题本身的难度

随训练程度加深，偏差降低，方差升高

Ch.3 线性模型

最小二乘法：略 P54-56 注意多元情形$\boldsymbol{X}^T\boldsymbol{X}$不满秩时：根据归纳偏好选择 P6；引入正则化 P129&P252

广义线性模型：$y=g^{-1}(\boldsymbol{w}^T\boldsymbol{x}+b)$,$g(\cdot)$单调可微，称为联系函数。$g(\cdot)=ln(\cdot)$以用于分类问题

对数几率回归：Sigmoid函数是单位阶跃函数的光滑近似，$y=\frac{1}{1+\exp(-(\boldsymbol{w}^T\boldsymbol{x}+b))}$

参数更新：极大似然法，对似然函数求极值，用梯度下降/牛顿法 P59-60

LDA：找到这样一条直线，使相同样例在其上的投影尽可能近，不同样例间尽可能远。

投影后同一样例内的协方差：$\boldsymbol{w}^T\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小

投影后不同样例间的均值向量距离：$\|\boldsymbol{w}^T\boldsymbol{\mu}_0-\boldsymbol{w}^T\boldsymbol{\mu}_1\|_2^2$尽可能大

优化目标$J(\boldsymbol{w})=\frac{\|\boldsymbol{w}^T\boldsymbol{\mu}_0-\boldsymbol{w}^T\boldsymbol{\mu}_1\|_2^2}{\boldsymbol{w}^T\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}}=\frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$,也称广义瑞利商

类内散度矩阵$\boldsymbol{S}_w=\Sigma_0+\Sigma_1$, 类间散度矩阵$\boldsymbol{S}_b=(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T$

优化过程见P61-62, 解得$\boldsymbol{w}=\boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$

LDA推广到多分类略，见P62-63

多分类学习：除了将二分类方法推广到多分类，更常用的是用拆分策略转化为二分类问题再集成

OvO：N个类别两两配对出$\frac{N(N-1)}{2}$个二分类器，最终的预测由投票得出

OvR：每个类别对其余所有类别，共N个二分类器，最终预测由预测置信度得出

对比：OvO存储开销和测试时间大，训练时间短；OvR相反，

预测性能取决于数据，多数情况差不多 P63-64

MvM：最常用 ECOC，用不同的正负例划分(列f_i)得到编码。 P64-65

优势：对分类器错误有一定的容忍修正能力，编码越长，编码距离越远，纠错能力越强。

类别不平衡：认为训练集是真实样本分布的无偏采样，将预测规则更改 P66-67

$\frac{y}{1-y}>1\Rightarrow\frac{y}{1-y}>\frac{m^+}{m^-}$

再缩放技术：欠采样(EasyEnsemble),过采样(SMOTE),阈值移动

欠采样丢弃大量样本，时间开销远小于增多样本的过采样

欠采样易丢失关键信息；过采样易造成过拟合

Ch.4 决策树 ⋆⋆

目的：产生一棵泛化能力强，也即处理未见样例能力强的决策树

划分选择： 希望分支结点包含的样本尽可能属于同一类别，也即"纯度"越来越高。

信息熵: 度量样本集合纯度常用指标，为$Ent(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k$

其中p_k为样本集合D中第k类样本所占比例，约定$p=0$时，$p\log_2 p=0$

信息熵最小值为0，最大值为$\log_2|\mathcal{Y}|$，二分类时$|\mathcal{Y}|=2$

信息增益： $Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$

表示用a进行划分产生V个分支结点，第v个分支上所有取值为a^v的样本记为D^v

选取信息增益最大的a

计算过程（ID3) 见课本P75-77

增益率： $Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)},IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$

信息增益的弊端：对分支数目较多的属性有偏好。

$IV(a)$称为属性a的固有值，a取值越多，该值越大。这可能对分支数目较少有偏好

C4.5算法使用了启发式的方法规避上述问题。P78-79

基尼值：$Gini(D)=\sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k}p_k p_{k'}=1-\sum_{k=1}^{|\mathcal{Y}|}p_k^2$

基尼值反映了从D中随机抽样两个样本类别不一致的概率。

基尼指数： $Gini\_index(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}Gini(D^v)$,选取基尼指数最小的$a$

CART算法的过程可见《统计学习方法》P84-85

剪枝处理： 决策树算法解决分支过多导致过拟合的主要方法，用留出法判断决策树泛化性能是否提升。

预剪枝： 生成过程中，每个结点按划分准则划分前先用验证集估计能否带来性能提升，

仅当验证集精度提升时才划分。

优点：降低过拟合风险，显著减少训练和测试的时间开销

缺点：预剪枝基于"贪心"算法的本质阻碍分支展开，存在欠拟合风险。

后剪枝： 对一颗完整的决策树，自底向上对非叶子结点考察，

若其对应的子树替换为叶子结点能使得验证集精度提升，则将其替换为叶子结点

优点：保留更多分支，欠拟合风险小，泛化性能一般比预剪枝好

缺点：自底向上对所有非叶子结点逐一考察，训练时间开销大

连续与缺失值：

连续值处理：连续属性离散化（二分法）实现过程见课本P84-85

缺失值处理：用不缺失的样本子集修改信息增益算式，见课本P86-88

多变量决策树： 属性a（非叶子节点）从单个属性变成多个属性的线性组合，略 P88-91

Ch.5 神经网络

定义：看课本P97-98

感知机：感知机由两层神经元组成，输入层接收信号，传递给输出层的M-P神经元，

一般输出记为$y=f(\sum_i w_i x_i-\theta)$,权重$w_i$和阈值$\theta$由学习得来

$w_i \leftarrow w_i+\Delta w_i$,其中$\Delta w_i=\eta(y-\hat{y})x_i$,$\eta\in(0,1)$为学习率

感知机只能解决线性可分问题，其在二维空间上的VC维仅为3。P98-100

多层前馈神经网络：非线性可分问题，输入和输出层间加隐层，隐层和输出层都是M-P神经元

神经网络的学习蕴含在连接权和阈值中

BP算法：手推过程见课本P102-104

标准BP算法：每次针对单个训练样例更新权值，更新频繁，多次迭代

累计BP算法：最小化训练集上的累计误差，读取整个训练集才更新参数，更新频率低

训练集非常大时，累计误差下降缓慢，标准BP算法的解较好

表示能力：隐层包含足够多神经元，多层网络能以任意精度逼近任意复杂度连续函数

局限性：1）易过拟合，通常用早停和正则化两种策略 P105

2）隐层神经元设置个数，通常用”试错法“调整

全局最小与局部极小值：下面的技术大多是启发式的，缺乏理论保障。

只谈策略：1）多组不同初始化，选loss最小的做最终参数；

2）模拟退火，每步以一定概率接受比当前更差的结果

3）SGD　4）遗传算法

其他常见神经网络：RBF网络，ART网络，SOM网络，级联相关网络，Elman网络，Boltzmann机

Ch.6 SVM

在样本空间上找一划分超平面以划分不同样本，位于"正中间"的超平面容忍性好，鲁棒性高，泛化能力强。

对一个能将所有样本类别正确划分的超平面，其线性方程为$\boldsymbol{w}^T\boldsymbol{x}+b=0$

空间中任意样本点$\boldsymbol{x}$到超平面距离可写为 $r=\frac{|\boldsymbol{w}^T\boldsymbol{x}+b|}{\|\boldsymbol{w}\|}$

因为超平面将所有类别正确划分，故$\boldsymbol{w}^T\boldsymbol{x}+b>0$当$y_i=1$, 反之亦然;

故可由伸缩变换$\boldsymbol{w}'=\zeta\boldsymbol{w}, b'=\zeta b$ 使$|\boldsymbol{w}^T\boldsymbol{x}_i+b|\geq 1, \forall x_i\in D$

此时距离超平面最近的异类支持向量到超平面间 "间隔" 为$\gamma=\frac{2}{\|\boldsymbol{w}\|}$

最大化间隔问题等效为：$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2, s.t.\, y_i(\boldsymbol{w}^T\boldsymbol{x_i}+b)\geq 1$

对偶问题：用Lagrange乘数法，添加乘子$\alpha_i\geq 0$, Lagrange函数为

$$L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x_i}+b))$$

$\boldsymbol{w}$和b求偏导取零，有$\boldsymbol{w}=\sum_{i=1}^m\alpha_i y_i\boldsymbol{x}_i, \sum_{i=1}^m\alpha_i y_i=0$代回上式即优化问题的对偶问题

$$\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^T\boldsymbol{x}_j, s.t.\sum_{i=1}^{m}\alpha_i y_i=0, \alpha_i\geq 0$$

该问题的优化采用**SMO算法**，首先注意到原优化问题存在不等式约束，迁移过来为**KKT条件**

$$\alpha_i\geq 0,\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1\geq 0,\ \alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1)=0$$

若$\alpha_i=0$，对结果无影响；若$y_i f(\boldsymbol{x}_i)-1=0$,样本点必位于最大间隔边界上(即支持向量)

解出最优的$\boldsymbol{\alpha}$即可代回求得$\boldsymbol{w}=\sum_{i=1}^m\alpha_i y_i\boldsymbol{x}_i, b=y_j-\sum_{i=1}^m\alpha_i y_i(\boldsymbol{x}_i^T\boldsymbol{x}_j))$

b中的(x_j,y_j)点要在最大间隔边界取到。得到的解仅与支持向量有关，具有稀疏性。

SMO算法：不断执行如下两步骤直至收敛：

1）选取一对更新变量$\alpha_i\&\alpha_j$

2）固定其他所有参数，求解对偶问题更新$\alpha_i\&\alpha_j$

此时约束$\sum_{i=1}^m\alpha_i y_i=0$变为$\alpha_i y_i+\alpha_j y_j=-\sum_{k\neq i,j}\alpha_k y_k$,解出一个变量代回

具体步骤可参考《统计学习方法》P124 例7.2，选择方法见同书的P147

核函数：不存在正确划分所有样本的超平面，故将其映射到更高维特征空间，使其线性可分。

将式子中样本$\boldsymbol{x}$变为映射后$\phi(\boldsymbol{x})$,

超平面为$f(\boldsymbol{x})=\boldsymbol{w}^T\phi(\boldsymbol{x})+b=\sum_{i=1}^m\alpha_i y_i\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x})+b$

求解过程中只涉及$\phi(\boldsymbol{x})$内积,用核函数$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$代替

常用核函数：P128

软间隔：核函数难以确定，引入软间隔，允许其在一些样本上出错，但这样的情况应尽可能少

优化目标 $\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^m l_{0/1}(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1)$

其中$l_{0/1}(z)=\begin{cases}1, z<0\\0, \text{otherwise}\end{cases}$称为"0/1损失函数",非凸非连续，三种替代见P130

软间隔对出错样本的容忍程度与C的取值有关

引入了松弛变量的”软间隔支持向量机“见P130-132

正则化：支持向量机模型的一般形式

$$\min_f \Omega(f)+C\sum_{i=1}^{m} l(f(\boldsymbol{x}_i),y_i)$$

$\Omega(f)$称为结构风险，描述模型自身性质；第二项称为经验风险，描述模型与数据契合程度。

参数C对二者进行折中，故也可理解为正则化问题的参数

SVR和**核方法**往年ppt没细讲，如有需要看P133-139

Ch.7 贝叶斯分类器 ⋆⋆

贝叶斯决策论：

考虑一多分类任务，所有类别为$\{c_1,\cdots,c_N\}$,λ_{ij}是将真实标记为c_j的样本误判为c_i产生的损失

基于后验概率$P(c_i|\boldsymbol{x})$可得到将样本$\boldsymbol{x}$分类为c_i产生的条件风险$R(c_i|\boldsymbol{x})=\sum_{j=1}^{N}\lambda_{ij}P(c_j|\boldsymbol{x})$

目标是找到一个分类器h以最小化总体期望风险$R(h)=\mathbb{E}_{\boldsymbol{x}}[R(h(\boldsymbol{x})|\boldsymbol{x})]$

贝叶斯判定准则$h^*(\boldsymbol{x})=\underset{c\in\mathcal{Y}}{\arg\min} R(c|\boldsymbol{x})$,贝叶斯最优分类器$h^*$，贝叶斯风险$R(h^*)$

$1-R(h^*)$反映了分类器能达到的最好性能

具体而言，只考虑最小化分类错误率，也即所有误判时损失都一致,$\lambda_{ij}=\begin{cases}0, i=j\\ 1,\text{otherwise}\end{cases}$

则条件风险为$R(c|\boldsymbol{x})=1-P(c|\boldsymbol{x})$,贝叶斯最优分类器为$h^*=\underset{c\in\mathcal{Y}}{\arg\min} P(c|\boldsymbol{x})$

即对每个样本$\boldsymbol{x}$选择使后验概率最大的类别。

问题是如何找后验概率$P(c|\boldsymbol{x})$,有两种策略：

判别式模型：直接建模$P(c|\boldsymbol{x})$预测c,如决策树，BP神经网络，支持向量机

生成式模型：对联合分布$P(\boldsymbol{x},c)$建模，用贝叶斯公式求得后验概率

由贝叶斯公式，$P(c|\boldsymbol{x})=\frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}$,其中$P(\boldsymbol{x})$为归一化因子，与类标记无关

问题转化为根据训练集D求解先验概率$P(c)$和似然$P(\boldsymbol{x}|c)$.

极大似然估计：都是统计学理论，很熟了，不用整理，需要看的话P149-150

朴素贝叶斯分类器：

提出”属性条件独立性假设“，也即每个特征对分类结果的影响相互独立，此时改写为

$$P(c|\boldsymbol{x})=\frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}=\frac{P(c)}{P(\boldsymbol{x})}\prod_{i=1}^{d}P(x_i|c)$$

其中d为特征数，x_i为样本$\boldsymbol{x}$在第i个特征上的取值。

此时有朴素贝叶斯分类器(Naïve Bayesian Classifier)

$$h_{nb}(\boldsymbol{x})=\underset{c\in\mathcal{Y}}{\arg\max} P(c)\prod_{i=1}^{d}P(x_i|c)$$

计算过程：先验概率

$P(Y=c_k)=\frac{\sum_{i=1}^{m}\mathbb{I}(y_i=c_k)}{m},P(\boldsymbol{x}^{(i)}=a_j|Y=c_k)=\frac{\sum_{i=1}^{m}\mathbb{I}(\boldsymbol{x}_i^{(j)}=a_j,y_i=c_k)}{\sum_{i=1}^{m}\mathbb{I}(y_i=c_k)}$

确定$\boldsymbol{x}$的类只需返回$P(Y=c_k)\prod_{j=1}^{m}P(X^{(j)}=\boldsymbol{x}^{(j)}|Y=c_k)$的最大值

例题步骤见《统计学习方法》P63 例4.1 和课本P151-153

拉普拉斯修正：训练集上从未与某类同时出现的单个属性值，在测试集上出现时似然必为0

做平滑修正$\hat{P}(c)=\frac{|D_c|+1}{|D|+N},\hat{P}(x_i|c)=\frac{|D_{c,x_i}|+1}{|D|+N_i}$

步骤见《统计学习方法》P64 例4.2 和课本P153-154

现实应用：若对预测速度要求高，可先将所有概率全计算出来，预测时直接查表。

若数据更替频繁，采用”懒惰学习“，需要预测时再计算概率

若数据不断增加，在现有估值基础上，对新增样本涉及的概率进行修正，增量学习

半朴素贝叶斯分类器：属性条件独立性假设在现实中很难达到

独依赖估计：假设每个属性在类别之外最多只依赖一个其他属性

$$P(c|\boldsymbol{x})\propto P(c)\prod_{i=1}^{d}P(\boldsymbol{x}_i|c,pa_i)$$

其中pa_i为属性$\boldsymbol{x}_i$所依赖的属性，称为$\boldsymbol{x}_i$的父属性。

各种独依赖分类器：SPODE，TAN，AODE P155-156

贝叶斯网：也称”信念网“，借助**有向无环图**刻画属性间依赖，使用**条件概率表**表述联合概率分布。P157

假设变为：每个属性与它的非子属性独立。见P157的示例

贝叶斯网中三个变量的典型依赖关系：同父结构，V型结构，顺序结构

如何分析有向图中变量的条件独立性？**有向分离！**

由有向图生成**道德图**：1）V型结构父属性相连 2）有向边变为无向边，具体分析见P159

学习过程：根据训练集找出结构最”恰当“的贝叶斯网

用评分函数评估”恰当“程度：**最小描述长度**,设训练集D,贝叶斯网$B=\langle G,\Theta\rangle$

$$S(B|D)=f(\theta)|B|-LL(B|D),\text{其中}LL(B|D)=\sum_{i=1}^{m}\log P_B(x_i)$$

$|B|$表示贝叶斯网参数个数，$f(\theta)$为描述每个参数θ所需编码位数

易见第一项表示贝叶斯网本身复杂度，第二项表示拟合程度,常用的有

$$AIC(B|D)=|B|-LL(B|D),BIC(B|D)=\frac{\log m}{2}|B|-LL(B|D)$$

推断过程：Gibbs采样 初始点q^0由已知变量观测值得到,每一步由随机游走得到

后验概率可近似为$P(Q=q|E=e)\simeq\frac{n_q}{T}$ 见P162

EM算法：针对数据有缺失的样本，未观测到的变量称为”隐变量“，模型参数Θ的对数似然

$$LL(\Theta|X,Z)=\ln P(X,Z|\Theta)$$

其中X表示观测到的变量集，Z表示隐变量集。Z无法得到，需计算期望得到边际似然

$$LL(\Theta|X) = \ln P(X|\Theta) = \ln \sum_Z P(X,Z|\Theta)$$

描述性步骤见课本P163,具体题目步骤见《统计学习方法》P175-177 例9.1

Ch.8 集成学习

目标：结合多个弱学习器以提升性能，每个个体应该”好而不同“

集成的示例：简单的二分类问题，每个基分类器的错误率$P(h_i(\boldsymbol{x}) \neq f(\boldsymbol{x})) = \varepsilon$

集成通过T个基分类器投票，以过半数为依据，$H(\boldsymbol{x}) = \mathrm{sign}(\sum_{i=1}^{T} h_i(\boldsymbol{x}))$

由Hoeffding ineq. $P(H(\boldsymbol{x}) \neq f(\boldsymbol{x})) \leq \exp\left(-\frac{1}{2}T(1-2\varepsilon)^2\right)$

则随着集成分类器的增加，错误率有指数级的下降趋于0

Caution:该分析基于各个基学习器的误差相互独立的假设，但这显然不可能

Boosting: 基学习器间存在强依赖关系，必须按序串行生成

机制：从初始训练集得到基学习器，提升基学习器判断错误样本的权重，重复训练

AdaBoost算法：做题步骤见《统计学习方法》P158-160

算法描述和理论推导见课本P174-176

Boosting要求在每轮训练时为每个样本采用”重赋权法“，

若基学习算法不接受权值，则可采用”重采样法“

Caution: 每轮训练后都要检查当前基学习器是否优于随机猜测，否则训练直接结束，

若采用”重采样”则可“重启动”避免训练过早停止

从偏差-方差角度理解，Boosting主要关注降低偏差，对泛化性能弱的学习器有很强集成

Bagging：基学习器不存在强依赖关系，可并行生成，基于第2章的自助采样法。

算法伪代码见课本P178，采样出T个采样集，基于每个采样集训练出基学习器再结合

Bagging对基学习器的结合通常用简单的投票法或平均法

优势：时间复杂度低，分析见课本P179; 从未被采样过的样本可作为验证集来“包外估计”

Bagging关注降低方差，在易受样本扰动学习器（不剪枝决策树，神经网络）上效果更佳

随机森林：Bagging的扩展变种,在以决策树为基学习器的Bagging集成上，引入了随机属性选择

决策树在选择划分属性时选择最优，

而随机森林则是随机取样出包含k个属性子集，再在其中选择最优

基学习器的多样性来自样本扰动和属性扰动，增大基学习器差异度，提升泛化性能

结合策略： 学习器结合从三个方面有好处 P181

平均法：简单平均，加权平均

投票法：绝对多数投票法，相对多数投票法，加权投票法

学习法：用新的学习器来学习基学习器的结合方法，Stacking算法 P184

多样性：

误差-分歧分解：对基学习器“好而不同”的理论分析，看课本P185-186

结论：基学习器的准确性越高，多样性越大，集成效果越好

多样性度量： 用于度量集成中基学习器的多样性 P186-188

对二分类问题，有分类器h_i和h_j的预测结果列联表,不同度量由该表给出

常见的有不合度量，相关系数，Q-统计量，κ-统计量

κ-误差图：横坐标为κ值，纵坐标为平均误差，

越高表示准确性越低，越靠右表示多样性越小。

多样性增强： 数据样本扰动，输入属性扰动，输出表示扰动，算法参数扰动 P188-190

数据样本扰动：采样，对"不稳定基学习器"很有效，如决策树，神经网络

稳定基学习器：线性学习器，SVM，Naïve Bayesian，kNN

输入属性扰动：随机子空间算法 P189

输出表示扰动：翻转法，输出调剂法，ECOC法

算法参数扰动：负相关法

不同多样性增强机制可同时使用

Ch.9 聚类 ⋆⋆

无监督学习中最重要的部分

目标：将数据集中的样本划分为多个一般不相交的簇。

既可作为单独的学习任务也可作为其他学习任务的前驱。

符号化表述见课本P197

性能度量： 也称"有效性指标"，希望"簇内相似度"高，"簇间相似度"低

外部指标：将结果与某参考模型比较   内部指标: 直接考察而不使用参考模型

常用的外部指标和内部指标见课本 P198-199

距离计算： 距离度量的性质 P199-200

常用距离 P200-201

注意区分不同类别属性上距离使用差异

有序属性：Minkowski， 无序属性：VDM，混合属性：结合，重要性：加权

原型聚类： 先对原型初始化，再迭代更新求解

k-means: 伪代码和例题见 P202-203, **直观步骤《统计学习方法》P265 例14.2**

学习向量量化(LVQ): 假设数据样本带有类被标记，利用监督信息辅助聚类 P205

高斯混合聚类：假设数据样本分布符合高斯混合分布,计算极大似然 P206-210

密度聚类： 假设聚类结构能通过样本分布的紧密程度来确定，主要考察样本间的可连接性

DBSCAN：基于"邻域"参数来刻画样本分布紧密程度 见P211-213

层次聚类： 在不同层次对数据集进行划分，以形成树形的聚类结构

可以"自底而上"（AGNES），也可以"自顶而下"(DIANA)。

AGNES: 见课本 P215-216 **例题见《统计学习方法》P262 例14.1**

Ch.10 降维与度量学习 ⋆⋆

kNN： 给定训练样本和距离度量，对测试样本，找到训练集中与其距离最近的k个样本

kNN是懒惰学习，训练开销为零，只在预测时进行处理；相对的是“急切学习”

1NN在二分类问题上的性能：P226 错误率不超过贝叶斯最优分类器的两倍！

低维嵌入： 维数灾难问题：“密采样”条件在高维情况下样本数要求过高。

利用数学变换，将高维的属性空间映到低维embedding，使样本密度大幅提升。

多维缩放(MDS)：可以使原始空间中样本间距离在低维空间中得以保持 P227-229

PCA：算法见P230-231，做题步骤见《统计学习方法》P314-3f15 例16.1 参考P310定义

核化线性降维(KPCA): 用高斯核投影，步骤见P233

流形学习： 流形局部具有欧氏空间的性质，

等度量映射(Isomap)：近邻点欧式距离保持，转换为最短路径问题 P234-235

局部线性嵌入(LLE): 使邻域内线性关系在降维后得到继续保持. P235-237

度量学习： 动机：不寻找合适的低维空间，而是直接学习距离度量。近邻成分学习(NCA) P238-239

Ch.11 特征选择与稀疏学习

对当前学习任务，特征有相关特征，无关特征，冗余特征（本节不考虑）

特征选择：从特征集合中选出任务相关特征子集，且保证重要特征不丢失

优点：减轻维度灾难，降低学习难度

朴素想法：遍历所有可能的子集，组合爆炸，不可行。

子集搜索与评价： 产生一初始候选子集，基于对该子集评价结果产生下一候选子集

贪心搜索：前向：逐渐增加特征，后向：从完整的减少特征，双向：增加相关减少无关

子集评价：用信息熵，考察特征子集的信息增益 P249

j将搜索机制与子集评价机制相结合，即得特征选择方法

过滤式选择： 先用特征选择过程过滤原始数据，再用过滤后的特征训练模型

特征选择和后续学习器无关

Relief方法 P249 其中相关统计量的确定 P250 其扩展变体Relief-F可应用多分类

包裹式选择： 把最终要使用的学习器的性能作为特征子集的评价准则

对比：从学习器性能上看，比过滤式好；但要多次训练学习器，计算开销大

LVW方法：P251

嵌入式选择： 将特征选择过程与学习器训练融为一体，在同一个优化过程中完成

考虑线性回归模型，引入L_2正则化项，有岭回归；替换为L_1范数，为LASSO回归

L_1范数有额外的好处，更易获得“稀疏”解，使求得的$\boldsymbol{w}$有更少的非零分量

L_1正则化问题的求解可采用近端梯度下降(PGD) P253-254

稀疏表示与字典学习： 稀疏矩阵中存在大量零元素，并非整行整列出现，其作为样本数据有如下优势:

1) 文本数据线性可分 2）存储高效

希望将稠密的数据集转化为”稀疏表示“，以享受上述优势。

为稠密的样本找到合适的**字典**而转化为稀疏表示，称为**字典学习**, 见P256

压缩感知： 利用部分数据恢复全部数据 P257-260

Ch.13 半监督学习

主动学习：学习获取到 “难” 分类预测样本，为其人工添加标签，将得到的样本加入训练集以提升性能

纯半监督学习：假定训练数据中未标记样本与待预测数据无关。

直推学习：假定学习过程中考虑的未标记样本恰是待预测数据。

要利用未标记样本，首先要明确一些假设

聚类假设：假设数据存在簇结构，同一簇的样本属于同一类别

流形假设：数据分布在一个流形结构上，邻近的样本具有相似的输出值

生成式方法： 假设所有数据由同一个潜在模型生成, 未标记数据视为模型的缺失参数 P296-297

高斯混合模型的参数估计可由EM算法求解。

优点：方法简单，易于实现，有标签数据量极少时性能更好

Caution: 模型假设必须准确，否则会显著降低泛化性能

半监督SVM（S3VM）: 试图找到能将两类有标记样本分开，且穿过数据低密度区域的划分超平面。

TSVM: 采用局部搜索来迭代寻找近似解，算法伪代码P300

标签指派和调整过程可能出现类别不平衡问题，可进行改进 P299-300

对每一对未标记样本进行调整，仍需要巨大的计算开销

需要设计出高效的优化求解策略，如

基于图核函数梯度下降的LaplacianSVM

基于标记均值估计的meanS3VM

图半监督学习： 将给定数据集映射为图，每个样本对应结点，若两样本相似度高，则存在一条边

边的强度正比于样本相似度

图对应其邻接矩阵，能够基于矩阵运算进行半监督学习 P301-304

缺点：存储开销大，且构图过程仅能考虑训练集，难以判断新样本在图中位置

基于分歧的方法： 使用多学习器,考察不同学习器间的”分歧“，重要代表为”协同训练 (co-training)“

过程：在每个视图上基于有标记样本训练出分类器；

让每个分类器挑选最有把握的未标记样本赋予伪标签；

将伪标签提供给另一分类器作为新增的有标记样本迭代更新。

若两个视图充分且条件独立，则可利用co-training将弱分类器的泛化性能提升到任意高

缺点：条件独立性难以满足，但即使在更弱的条件下，co-training仍可提升弱分类器性能

当有标记样本很少、尤其是数据不具有多视图时，很难生成具有显著分歧的学习器

半监督聚类：利用监督信息以获得更好的聚类效果

监督信息类型：1）“ 必连”与“勿连”: 前者表示样本必属于同一簇，后者表示必不属于同一簇

2）少量的有标记样本

约束*k*均值算法：利用第一类监督信息，必须确保该类信息约束满足 P307-309

约束种子*k*均值：直接将监督信息作为“种子”, 用其初始化*k*均值算法的*k*个聚类中心,

并且在聚类迭代更新过程中不改变种子样本所属的簇. P309-310

讨论：半监督学习在利用未标记样本后并非必然提升泛化性能, 在有些情形下甚至会导致性能下降.

生成式方法: 模型假设不准确, 需依赖充分可靠的领域知识来设计模型

半监督SVM: 训练数据中存在多个“低密度划分”, 算法有可能做出不利的选择

S4VM优化最坏情形性能来综合利用多个低密度划分, 提升了此类技术的安全性